CS 513 Knowledge Disc and Data Mining

Mid Term

#1 (10 Points)

Is the following function a proper distance function? Why? Explain your answer.

$$d(\mathbf{x},\mathbf{y}) = (\sum_i | x_i - y_i |)^3$$

Let us assume that, X = (0,0), Y = (0,1) and Z = (1,1)

Using given distance function,

The distance between X (0,0) & Y (0,1) => d (x, y)

$= (|0 – 0| + |0 – 1|)^3$
$= (0+ 1)^3$
$= (1)^3$
$= 1$

The distance between Y (0,1) & X (0,0) => d (y, x)

$= (|0 – 0| + |1 – 0|)^3$
$= (0 + 1)^3$
= (1)3
= 1

The distance between Y (0,1) & Z (1,1) => d (y, z)

$= (|0 – 1| + |1 – 1|)^3$
$= (1 + 0)^3$
$= (1)^3$
= 1

The distance between Z (1,1) & Y (0,1) => d (z, y)

$= (|1 – 0| + |1 – 1|)^3$
$= (1+ 0)^3$
$= (1)^3$
= 1

The distance between Z (1,1) & X (0,0) => d (z, x)

$= (|1 – 0| + |1 – 0|)^3$
$= (1 + 1)^3$
= (2)3
= 8

The distance between X (0,0) & Z (1,1) => d (x, z)

$= (|0 – 1| + |0 – 1|)^3$
$= (1 + 1)^3$

$= (2)^3$

$= 8$

Checking validity of the distance function properties on the distance values calculated using given distance function.

1. d (x, y) ≥ 0, d (y, x) ≥ 0, d (y, z) ≥ 0, d (z, y) ≥ 0, d (z, x) ≥ 0, d (x, z) ≥ 0.
   Clearly $d(x,y) \geq 0$ and $d(x,y) = 0 \Leftrightarrow x = y$ are satisfied.

2. d (x, y) = d (y, x), d (y, z) = d (z, y), d (z, x) = d (x, z)
   Clearly $d(x,y) = d(y,x)$ is satisfied.

3. d (x, z) = 8, d (x, y) = 1, d (y, z) = 1
   $d(x,z) \leq d(x,y) + d(y,z)$
   8 ≤ 1 + 1
   8 ≤ 2 which is false. So, condition 4 failed d (z, x)
   = 8, d (z, y) = 1, d (y, x) = 1.
   $d(z,x) \leq d(z,y) + d(y,x)$
   8 ≤ 1 + 1
   8 ≤ 2 which is false. So, condition 4 failed here as well.

As per above calculations and observations, given distance function satisfies the first 3 conditions but fails to meet the last condition (Triangle inequality). Therefore, given function is not a proper distance function.

 # 2 (10 Points)

There are three major manufacturing companies that make a product: Manufacturers A, B, and C. Manufacture A has a 50% market share, and Manufacture B has a 30% market share. 5% of A's products are defective, 6% of B's products are defective, and 8% of C's products are defective.

a) What is the probability that a randomly selected product is defective? P(Defective)?
b) What is the probability that a randomly selected product is defective and that it came from A? P(A and Defective)?
c) What is the probability that a defective product came from B? P(B/Defective)?
d) Are these events (being defective and coming from B) independent? Why?

Solution:

Let's assume there are 1000 items of the product in the market => N = 1000

Based on Market Share,

A has 50% of market share. => N(A) = 50% of 1000 = 500

B has 30% of market share. => N(B) = 30% of 1000 = 300

Remaining are from C => N(c) = 1000-500-300 = 200

Number of defective pieces by manufacturer are as follows:

A's defective products = N (Defective | A) = 5% of 500 items = 25

B's defective products = N (Defective | B) = 6% of 300 items = 18

C's defective products = N (Defective | C) = 8% of 200 items = 16

a) P(Defective) = (N (Defective | A) + N (Defective | B) + N (Defective | C)) / N

= (25 + 18 + 16) / 1000 = 59 / 1000 = 0.059 = 5.9%

b) P (A ∩ Defective) = N (Defective | A) / N = 25 / 1000 = 0.025 = 2.5%

c) P (B | Defective) = P (Defective | B) / P(Defective) = 18 / 59 = 0.3051 = 30.51% **d)** P(B) = 300 / 1000 = 0.3

P(Defective) = 59 / 1000 = 0.059

For events to be independent => P (B ∩ Defective) = P(B) * P(Defective)

P(B) * P(Defective) = 0.3 * 0.059 = 0.0177

P (B ∩ Defective) = 18 / 1000 = 0.018

Since, P (B ∩ Defective) ≠ P(B) * P(Defective)

Therefore, the events are **not independent** of each other.